

# Low-distortion Subspace Embeddings in Input-sparsity Time and Applications to Robust Linear Regression

Xiangrui Meng \*

Michael W. Mahoney †

## Abstract

Low-distortion subspace embeddings are critical building blocks for developing improved random sampling and random projection algorithms for common linear algebra problems. Here, we show that, given a matrix  $A \in \mathbb{R}^{n \times d}$ , with  $n \gg d$ , and a  $p \in [1, 2)$ , with a constant probability, we can construct a low-distortion embedding matrix  $\Pi \in \mathbb{R}^{\text{poly}(d) \times d}$  that embeds  $\mathcal{A}_p$ , the  $\ell_p$  subspace spanned by  $A$ 's columns, into  $(\mathbb{R}^{\mathcal{O}(\text{poly}(d))}, \|\cdot\|_p)$ ; the distortion of our embeddings is only  $\mathcal{O}(\text{poly}(d))$ , and we can compute  $\Pi A$  in  $\mathcal{O}(\text{nnz}(A))$  time, i.e., input-sparsity time. Our result generalizes the input-sparsity time  $\ell_2$  subspace embedding proposed recently by Clarkson and Woodruff; and for completeness, we present a simpler and improved analysis of their construction for  $\ell_2$ . These input-sparsity time  $\ell_p$  embeddings are optimal, up to constants, in terms of their running time; and the improved running time propagates to applications such as  $(1 \pm \epsilon)$ -distortion  $\ell_p$  subspace embedding and relative-error  $\ell_p$  regression. For  $\ell_2$ , we show that a  $(1 + \epsilon)$ -approximate solution to the  $\ell_2$  regression problem specified by the matrix  $A$  and a vector  $b \in \mathbb{R}^n$  can be computed in  $\mathcal{O}(\text{nnz}(A) + d^5 \log(d/\epsilon)/\epsilon^2)$  time; and for  $\ell_p$ , via a subspace-preserving sampling procedure, we show that a  $(1 \pm \epsilon)$ -distortion embedding of  $\mathcal{A}_p$  into  $\mathbb{R}^{\mathcal{O}(\text{poly}(d))}$  can be computed in  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time, and we also show that a  $(1 + \epsilon)$ -approximate solution to the  $\ell_p$  regression problem  $\min_{x \in \mathbb{R}^d} \|Ax - b\|_p$  can be computed in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) \log(1/\epsilon)/\epsilon^2)$  time. Moreover, we can also improve the embedding dimension or equivalently the sample size to  $\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2)$  without increasing the complexity.

## 1 Introduction

Regression problems are ubiquitous, and the fast computation of their solutions is of interest in many large-scale data applications. A parameterized family of regression problems that is of particular interest is the *overconstrained  $\ell_p$  regression problem*: given a matrix  $A \in \mathbb{R}^{n \times d}$ , with  $n > d$ , a vector  $b \in \mathbb{R}^n$ , a norm  $\|\cdot\|_p$  parameterized by  $p \in [1, \infty]$ , and an error parameter  $\epsilon > 0$ , find a  $(1 + \epsilon)$ -approximate solution  $\hat{x} \in \mathbb{R}^d$  to:

$$f^* = \min_{x \in \mathbb{R}^d} \|Ax - b\|_p, \quad (1)$$

i.e., find a vector  $\hat{x}$  such that  $\|A\hat{x} - b\|_p \leq (1 + \epsilon)f^*$ , where the  $\ell_p$  norm of a vector  $x$  is  $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ , defined to be  $\max_i |x_i|$  for  $p = \infty$ . Special cases include the  $\ell_2$  regression problem, also known as Least Squares Approximation problem, and the  $\ell_1$  regression problem, also known as the Least Absolute Deviations or Least Absolute Errors problem. The latter is of particular interest as a *robust estimation* or *robust regression* technique, in that it is less sensitive to the presence of outliers than the former. We are most interested in this paper in the  $\ell_1$  regression problem due to

\*ICME, Stanford University, Stanford, CA 94305. Email: mengxr@stanford.edu

†Dept. of Mathematics, Stanford University, Stanford, CA 94305. Email: mmahoney@cs.stanford.edu

its robustness properties, but our methods hold for general  $p \in [1, 2]$ , and thus we formulate our results in  $\ell_p$ .

It is well-known that for  $p \geq 1$ , the overconstrained  $\ell_p$  regression problem is a convex optimization problem; for  $p = 1$  and  $p = \infty$ , it is an instance of linear programming; and for  $p = 2$ , it can be solved with eigenvector-based methods such as with the QR decomposition or the Singular Value Decomposition of  $A$ . In spite of their low-degree polynomial-time solvability,  $\ell_p$  regression problems have been the focus in recent years of a wide range of random sampling and random projection algorithms, largely due to a desire to develop improved algorithms for large-scale data applications [3, 23, 10]. For example, Clarkson [9] uses subgradient and sampling methods to compute an approximate solution to the overconstrained  $\ell_1$  regression problem in roughly  $\mathcal{O}(nd^5 \log n)$  time; and Dasgupta et al. [12] use well-conditioned bases and subspace-preserving sampling algorithms to solve general  $\ell_p$  regression problems, for  $p \in [1, \infty)$ , in roughly  $\mathcal{O}(nd^5 \log n)$  time. A similar subspace-preserving sampling algorithm was developed by Drineas, Mahoney, and Muthukrishnan [16] to compute an approximate solution to the  $\ell_2$  regression problem. The algorithm of [16] relies on the estimation of the  $\ell_2$  leverage scores<sup>1</sup> of  $A$  to be used as an importance sampling distribution, but when combined with the results of Sarlós [27] and Drineas et al. [17] (that quickly preprocess  $A$  to uniformize those scores) or Drineas et al. [15] (that quickly computes approximations to those scores), this leads to a random projection or random sampling (respectively) algorithm for the  $\ell_2$  regression problem that runs in roughly  $\mathcal{O}(nd \log d)$  time [17, 20]. More recently, Sohler and Woodruff [28] introduced the Cauchy Transform to obtain improved  $\ell_1$  embeddings, thereby leading to an algorithm for the  $\ell_1$  regression problem that runs in  $\mathcal{O}(nd^{1.376+})$  time; and Clarkson et al. [10] use the Fast Cauchy Transform and ellipsoidal rounding methods to compute an approximation to the solution of general  $\ell_p$  regression problems in roughly  $\mathcal{O}(nd \log n)$  time.

These algorithms, and in particular the algorithms for  $p = 2$ , form the basis for much of the large body of recent work in randomized algorithms for low-rank matrix approximation, and thus optimizing their properties can have immediate practical benefits. See, e.g., the recent monograph of Mahoney [20] and references therein for details. Although some of these algorithms are near-optimal for dense inputs, they all require  $\Omega(nd \log d)$  time, which can be large if the input matrix is very sparse. Thus, it was a significant result when Clarkson and Woodruff [11] developed an algorithm for the  $\ell_2$  regression problem (as well as the related problems of low-rank matrix approximation and  $\ell_2$  leverage score approximation) that runs in *input-sparsity time*, i.e., in  $\mathcal{O}(\text{nnz}(A) + \text{poly}(d/\epsilon))$  time, where  $\text{nnz}(A)$  is the number of non-zero elements in  $A$  and  $\epsilon$  is an error parameter. This result depends on the construction of a *sparse embedding matrix*  $\Pi$  for  $\ell_2$ . By this, we mean the following: for an  $n \times d$  matrix  $A$ , an  $s \times n$  matrix  $\Pi$  such that,

$$(1 - \epsilon)\|Ax\|_2 \leq \|\Pi Ax\|_2 \leq (1 + \epsilon)\|Ax\|_2,$$

for all  $x \in \mathbb{R}^d$ . That is,  $\Pi$  embeds the column space of  $A$  into  $\mathbb{R}^s$ , while approximately preserving the  $\ell_2$  norms of all vectors in that subspace. Clarkson and Woodruff achieve their improved results for  $\ell_2$ -based problems by showing how to construct such a  $\Pi$  with  $s = \text{poly}(d/\epsilon)$  and showing that it can be applied to an arbitrary  $A$  in  $\mathcal{O}(\text{nnz}(A))$  time [11]. (In particular, this embedding result improves the result of Meng, Saunders, and Mahoney [23], who in their development of the parallel least-squares solver LSRN use a result from Davidson and Szarek [14] to construct a

---

<sup>1</sup>Recall that for an  $n \times d$  matrix  $A$ , with  $n \gg d$ , the  $\ell_2$  leverage scores of the rows of  $A$  are equal to the diagonal elements of the projection matrix onto the span of  $A$ . That is, if  $A = QR$  is a QR decomposition of  $A$ , or if  $A = Q\Sigma V^T$  is the thin SVD of  $A$ , then the leverage scores equal the Euclidean norms squared of the rows of the  $n \times d$  matrix  $Q$ , and thus they can be computed exactly in  $\mathcal{O}(nd^2)$  time. See [20, 15] for details; and note that they can be generalized to  $\ell_1$  and other  $\ell_p$  norms [10] as well as to arbitrary  $n \times d$  matrices, with both  $n$  and  $d$  large, if one specifies a low-rank parameter [21, 15].

constant-distortion embedding for  $\ell_2$  that runs in  $\mathcal{O}(\text{nnz}(A) \cdot d)$  time.) Interestingly, the analysis of Clarkson and Woodruff coupled ideas from the data streaming literature with the structural fact that there cannot be too many high-leverage constraints/rows in  $A$ . In particular, they showed that the high-leverage parts of the subspace may be viewed as heavy-hitters that are “perfectly hashed,” and thus contribute no distortion, and that the distortion of the rest of the subspace as well as the “cross terms” may be bounded with a result of Dasgupta, Kumar, and Sarlós [13].

In this paper, we provide improved low-distortion subspace embeddings for  $\ell_p$ , for all  $p \in [1, 2]$ , in input-sparsity time; and we show that, by coupling with recent work on fast subspace-preserving sampling from [10], these embeddings can be used to provide  $(1 + \epsilon)$ -approximate solutions to  $\ell_p$  regression problems, for  $p \in [1, 2]$ , in nearly input-sparsity time. In more detail, our main results are the following.

- For  $\ell_2$ , we obtain an improved result for the input-sparsity time  $(1 \pm \epsilon)$ -distortion embedding of [11]. In particular, for the same embedding procedure, we obtain improved bounds for the embedding dimension with a much simpler analysis than [11]. See Theorem 1 of Section 3 for a precise statement of this result. Our analysis is direct and does *not* rely on splitting the high-dimensional space into a set of heavy-hitters consisting of the high-leverage components and the complement of that heavy-hitting set. In addition, since our result directly improves the  $\ell_2$  embedding result of Clarkson and Woodruff [11], it immediately leads to improvements for the  $\ell_2$  regression, low-rank matrix approximation, and  $\ell_2$  leverage score estimation problems that they consider.
- For  $\ell_1$ , we obtain a low-distortion sparse embedding matrix  $\Pi$  such that  $\Pi A$  can be computed in input-sparsity time. That is, we construct an embedding matrix  $\Pi \in \mathbb{R}^{\text{poly}(d) \times n}$  such that, for all  $x \in \mathbb{R}^d$ ,

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_1 \leq \|\Pi Ax\|_1 \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_1,$$

with a constant probability, and  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A))$  time. See Theorem 2 of Section 4 for a precise statement of this result. Here, our proof involves splitting the set  $Y = \{Ux \mid \|x\|_\infty = 1, x \in \mathbb{R}^d\}$ , where  $U$  is an  $\ell_1$  well-conditioned basis for the span of  $A$ , into two parts, informally a subset where coordinates of high  $\ell_1$  leverage dominate  $\|y\|_1$  and the complement of that subset. This  $\ell_1$  result leads to immediate improvements in  $\ell_1$ -based problems. For example, by taking advantage of the fast version of subspace-preserving sampling from [10], we can construct and apply a  $(1 \pm \epsilon)$ -distortion sparse embedding matrix for  $\ell_1$  in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$  time. In addition, we can use it to compute a  $(1 + \epsilon)$ -approximation to the  $\ell_1$  regression problem in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$  time, which in turn leads to immediate improvements in  $\ell_1$ -based matrix approximation objectives, e.g., for the  $\ell_1$  subspace approximation problem [6, 28, 10].

- For  $\ell_p$ , for all  $p \in (1, 2)$ , we obtain a low-distortion sparse embedding matrix  $\Pi$  such that  $\Pi A$  can be computed in input-sparsity time. That is, we construct an embedding matrix  $\Pi \in \mathbb{R}^{\text{poly}(d) \times n}$  such that, for all  $x \in \mathbb{R}^d$ ,

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p,$$

with a constant probability, and  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A))$  time. See Theorem 4 of Section 5 for a precise statement of this result. Here, our proof generalizes the  $\ell_1$  result, but we need to prove upper and lower tail bound inequalities for sampling from general  $p$ -stable distributions that are of independent interest. Although these distributions don’t

have closed forms for  $p \in (1, 2)$  in general, we prove that there exists an order among the Cauchy distribution, a  $p$ -stable distribution with  $p \in (1, 2)$ , and the Gaussian distribution such that for all  $p \in (1, 2)$  we can use the upper bound from the Cauchy distribution and the lower bound from the Gaussian distribution. As with our  $\ell_1$  result, this  $\ell_p$  result has several extensions: in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$  time, we can construct and apply a  $(1 \pm \epsilon)$ -distortion sparse embedding matrix for  $\ell_p$ ; in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d/\epsilon))$  time, we can compute a  $(1 + \epsilon)$ -approximation to the  $\ell_p$  regression problem; and in  $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$  time, we can construct and apply a near-optimal (in terms of embedding dimension and distortion factor) embedding matrix.

The  $(1 \pm \epsilon)$ -distortion subspace embedding (for  $\ell_p$ ,  $p \in [1, 2]$ , that we construct from the input-sparsity time embedding and the fast subspace-preserving sampling) has embedding dimension  $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ , where the somewhat large  $\text{poly}(d)$  term directly multiplies the  $\log(1/\epsilon)/\epsilon^2$  term. We can also improve this, showing that it is possible, without increasing the overall complexity, to decouple the large  $\text{poly}(d)$  and  $\log(1/\epsilon)/\epsilon^2$  via another round of sampling and conditioning, thereby obtaining an embedding dimension that is a small  $\text{poly}(d)$  times  $\log(1/\epsilon)/\epsilon^2$ . See Theorem 7 of Section 6 for a precise statement of this result.

## 2 Background

We use  $\|\cdot\|_p$  to denote the  $\ell_p$  norm of a vector,  $\|\cdot\|_2$  the spectral norm of a matrix, and  $|\cdot|_p$  the element-wise  $\ell_p$  norm of a matrix. Given  $A \in \mathbb{R}^{n \times d}$  with full column rank and  $p \in [1, 2]$ , we use  $\mathcal{A}_p$  to denote the  $\ell_p$  subspace spanned by  $A$ 's columns. In this paper, we are interested in fast embedding of  $\mathcal{A}_p$  into a  $d$ -dimensional subspace of  $(\mathbb{R}^{\text{poly}(d)}, \|\cdot\|_p)$ , with distortion either  $\text{poly}(d)$  or  $(1 \pm \epsilon)$ , for some  $\epsilon > 0$ , as well as applications of this embedding to problems such as  $\ell_p$  regression. We assume that  $n \gg \text{poly}(d) \geq d \gg \log n$ . To state our results, we assume that we are capable of computing a  $(1 + \epsilon)$ -approximate solution to an  $\ell_p$  regression problem of size  $n' \times d$  for some  $\epsilon > 0$ , as long as  $n'$  is independent of  $n$ . Let us denote the running time needed to solve this smaller problem by  $\mathcal{T}_p(\epsilon; n', d)$ . In theory, we have  $\mathcal{T}_2(\epsilon; n', d) = \mathcal{O}(n'd \log(d/\epsilon) + d^3)$  (see Rokhlin and Tygert [26] and Drineas et al. [17]), and  $\mathcal{T}_p(\epsilon; n', d) = \mathcal{O}((n'd^2 + \text{poly}(d)) \log(n'/\epsilon))$ , for general  $p$  (see, e.g., Mitchell [24]).

**Conditioning.** The  $\ell_p$  subspace embedding and  $\ell_p$  regression problems are closely related to the concept of conditioning. We state here two related notions of  $\ell_p$ -norm conditioning and then a lemma that characterizes the relationship between them.

**Definition 1** ( $\ell_p$ -norm Conditioning (from [10])). *Given an  $n \times d$  matrix  $A$  and  $p \in [1, \infty]$ , let*

$$\sigma_p^{\max}(A) = \max_{\|x\|_2 \leq 1} \|Ax\|_p \text{ and } \sigma_p^{\min}(A) = \min_{\|x\|_2 \geq 1} \|Ax\|_p.$$

*Then, we denote by  $\kappa_p(A)$  the  $\ell_p$ -norm condition number of  $A$ , defined to be:*

$$\kappa_p(A) = \sigma_p^{\max}(A) / \sigma_p^{\min}(A).$$

*For simplicity, we will use  $\kappa_p$ ,  $\sigma_p^{\min}$ , and  $\sigma_p^{\max}$  when the underlying matrix is clear.*

**Definition 2** ( $(\alpha, \beta, p)$ -conditioning (from [12])). *Given an  $n \times d$  matrix  $A$  and  $p \in [1, \infty]$ , let  $q$  be the dual norm of  $p$ . Then  $A$  is  $(\alpha, \beta, p)$ -conditioned if (1)  $|A|_p \leq \alpha$ , and (2) for all  $z \in \mathbb{R}^d$ ,  $\|z\|_q \leq \beta \|Az\|_p$ . Define  $\bar{\kappa}_p(A)$  as the minimum value of  $\alpha\beta$  such that  $A$  is  $(\alpha, \beta, p)$ -conditioned.*

**Lemma 1** (Equivalence of  $\kappa_p$  and  $\bar{\kappa}_p$  (from [10])). *Given an  $n \times d$  matrix  $A$  and  $p \in [1, \infty]$ , we always have*

$$d^{-|1/2-1/p|} \kappa_p(A) \leq \bar{\kappa}_p(A) \leq d^{\max\{1/2, 1/p\}} \kappa_p(A).$$

**Remark.** Given the equivalence established by Lemma 1, we will say that  $A$  is *well-conditioned* in the  $\ell_p$  norm if  $\kappa_p(A)$  or  $\bar{\kappa}_p(A) = \mathcal{O}(\text{poly}(d))$ , independent of  $n$ .

Although for an arbitrary matrix  $A \in \mathbb{R}^{n \times d}$ , the condition numbers  $\kappa_p(A)$  and  $\bar{\kappa}_p(A)$  can be arbitrarily large, we can often find a matrix  $R \in \mathbb{R}^{d \times d}$  such that  $AR^{-1}$  is well-conditioned. This procedure is called *conditioning*, and there exist two approaches for conditioning: via low-distortion  $\ell_p$  subspace embedding and via ellipsoidal rounding.

**Definition 3** (Low-distortion  $\ell_p$  Subspace Embedding). *Given an  $n \times d$  matrix  $A$  and  $p \in [1, \infty]$ ,  $\Pi \in \mathbb{R}^{s \times n}$  is a low-distortion embedding of  $\mathcal{A}_p$  if  $s = \mathcal{O}(\text{poly}(d))$  and*

$$1/\mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|Ax\|_p, \quad \forall x \in \mathbb{R}^d.$$

**Remark.** Given a low-distortion embedding matrix  $\Pi$  of  $\mathcal{A}_p$ , let  $R$  be the “R” matrix from the QR decomposition of  $\Pi A$ . Then, the matrix  $AR^{-1}$  is well-conditioned in the  $\ell_p$  norm. To see this, note that we have

$$\|AR^{-1}x\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|\Pi AR^{-1}x\|_p \leq \mathcal{O}(\text{poly}(d)) \cdot \|\Pi AR^{-1}\|_2 = \mathcal{O}(\text{poly}(d)) \cdot \|x\|_2, \quad \forall x \in \mathbb{R}^d,$$

where the first inequality is due to low distortion and the second inequality is due to  $s = \mathcal{O}(\text{poly}(d))$ . By similar arguments, we can show that  $\|AR^{-1}x\|_p \geq 1/\mathcal{O}(\text{poly}(d)) \cdot \|x\|_2$ ,  $\forall x \in \mathbb{R}^d$ . Hence, by combining these results, the matrix  $AR^{-1}$  is well-conditioned in the  $\ell_p$  norm.

For a discussion of ellipsoidal rounding, we refer readers to Clarkson et al. [10]. In this paper, we simply cite the following lemma, which is based on ellipsoidal rounding.

**Lemma 2** (Fast  $\mathcal{O}(d)$ -conditioning (from [10])). *Given an  $n \times d$  matrix  $A$  and  $p \in [1, \infty]$ , it takes at most  $\mathcal{O}(nd^3 \log n)$  time to find a matrix  $R \in \mathbb{R}^{d \times d}$  such that  $\kappa_p(AR^{-1}) \leq 2d$ .*

**Subspace-preserving sampling and  $\ell_p$  regression.** Given  $R \in \mathbb{R}^{d \times d}$  such that  $AR^{-1}$  is well-conditioned in the  $\ell_p$  norm, we can construct a  $(1 \pm \epsilon)$ -distortion embedding, specifically a subspace-preserving sampling, of  $\mathcal{A}_p$  in  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  additional time and with a constant probability. This result from Clarkson et al. [10, Theorem 5.4] improves the subspace-preserving sampling algorithm proposed by Dasgupta et al. [12] by estimating the row norms of  $AR^{-1}$  (instead of computing them exactly) to define importance sampling probabilities.

**Lemma 3** (Fast Subspace-preserving Sampling (from [10])). *Given a matrix  $A \in \mathbb{R}^{n \times d}$ ,  $p \in [1, \infty]$ ,  $\epsilon > 0$ , and a matrix  $R \in \mathbb{R}^{d \times d}$  such that  $AR^{-1}$  is well-conditioned, it takes  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time to compute a sampling matrix  $S \in \mathbb{R}^{s \times n}$  (with only one nonzero element per row) with  $s = \mathcal{O}(\bar{\kappa}_p^p(AR^{-1})d^{|p/2-1|+1} \log(1/\epsilon)/\epsilon^2)$  such that with a constant probability,*

$$(1 - \epsilon)\|Ax\|_p \leq \|SAx\|_p \leq (1 + \epsilon)\|Ax\|_p, \quad \forall x \in \mathbb{R}^d.$$

Given such a subspace-preserving sampling algorithm, Clarkson et al. [10, Theorem 5.4] show that it is straightforward to compute a  $\frac{1+\epsilon}{1-\epsilon}$ -approximate solution to an  $\ell_p$  regression problem.

**Lemma 4** ( $\ell_p$  Regression via Sampling (from [10])). *Given an  $\ell_p$  regression problem specified by  $A \in \mathbb{R}^{n \times d}$ ,  $b \in \mathbb{R}^n$ , and  $p \in [1, \infty]$ , let  $S$  be a  $(1 \pm \epsilon)$ -distortion embedding matrix of the subspace spanned by  $A$ ’s columns and  $b$  from Lemma 3, and let  $\hat{x}$  be an optimal solution to the subsampled problem  $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_p$ . Then  $\hat{x}$  is a  $\frac{1+\epsilon}{1-\epsilon}$ -approximate solution to the original problem.*

**Remark.** Collecting these results, we see that a low-distortion  $\ell_p$  subspace embedding is a fundamental building block (and very likely a bottleneck) for  $(1 \pm \epsilon)$ -distortion  $\ell_p$  subspace embeddings, as well as for a  $(1 + \epsilon)$ -approximation to an  $\ell_p$  regression problem. This motivates our work and its emphasis on finding low-distortion subspace embeddings more efficiently.

**Stable distributions.** The properties of  $p$ -stable distributions are essential for constructing input-sparsity time low-distortion  $\ell_p$  subspace embeddings.

**Definition 4** ( $p$ -stable Distribution). *A distribution  $\mathcal{D}$  over  $\mathbb{R}$  is called  $p$ -stable, if for any  $m$  real numbers  $a_1, \dots, a_m$ , we have*

$$\sum_{i=1}^m a_i X_i \simeq \left( \sum_{i=1}^m |a_i|^p \right)^{1/p} X,$$

where  $X_i \stackrel{iid}{\sim} \mathcal{D}$  and  $X \sim \mathcal{D}$ . By “ $X \simeq Y$ ”, we mean  $X$  and  $Y$  have the same distribution.

By a result due to Lévy [19], it is known that  $p$ -stable distributions exist for  $p \in (0, 2]$ ; and from Chambers et al. [7], it is known that  $p$ -stable random variables can be generated efficiently, thus allowing their practical use. Let us use  $\mathcal{D}_p$  to denote the “standard”  $p$ -stable distribution, for  $p \in [1, 2]$ , specified by its characteristic function  $\psi(t) = e^{-|t|^p}$ . It is known that  $\mathcal{D}_1$  is the standard Cauchy distribution, and that  $\mathcal{D}_2$  is the Gaussian distribution with mean 0 and variance 2.

**Tail inequalities.** We note two inequalities from Clarkson et al. [10] regarding the tails of the Cauchy distribution.

**Lemma 5** (Cauchy Upper Tail Inequality). *For  $i = 1, \dots, m$ , let  $C_i$  be  $m$  (not necessarily independent) standard Cauchy variables, and  $\gamma_i > 0$  with  $\gamma = \sum_i \gamma_i$ . Let  $X = \sum_i \gamma_i |C_i|$ . For any  $t > 1$ ,*

$$\Pr[X > t\gamma] \leq \frac{1}{\pi t} \left( \frac{\log(1 + (2mt)^2)}{1 - 1/(\pi t)} + 1 \right).$$

For simplicity, we assume that  $m \geq 3$  and  $t \geq 1$ , and then we have  $\Pr[X > t\gamma] \leq 2 \log(mt)/t$ .

**Lemma 6** (Cauchy Lower Tail Inequality). *For  $i = 1, \dots, m$ , let  $C_i$  be independent standard Cauchy random variables, and  $\gamma_i \geq 0$  with  $\gamma = \sum_i \gamma_i$ . Let  $X = \sum_i \gamma_i |C_i|$ . Then, for any  $t > 0$ ,*

$$\log \Pr[X \leq (1 - t)\gamma] \leq \frac{-\gamma t^2}{3 \max_i \gamma_i}.$$

We also note the following result about Gaussian variables. This is a direct consequence of Maurer’s inequality ([22]), and we will use it to derive lower tail inequalities for  $p$ -stable distributions.

**Lemma 7** (Gaussian Lower Tail Inequality). *For  $i = 1, \dots, m$ , let  $G_i$  be independent standard Gaussian random variables, and  $\gamma_i \geq 0$  with  $\gamma = \sum_i \gamma_i$ . Let  $X = \sum_i \gamma_i |G_i|^2$ . Then, for any  $t > 0$ ,*

$$\log \Pr[X \leq (1 - t)\gamma] \leq \frac{-\gamma t^2}{6 \max_i \gamma_i}.$$

### 3 Main Results for $\ell_2$ Embedding

Here is our main result for input-sparsity time low-distortion subspace embeddings for  $\ell_2$ .

**Theorem 1** ( $(1 \pm \epsilon)$ -distortion Embedding for  $\ell_2$ ). *Given a matrix  $A \in \mathbb{R}^{n \times d}$  and  $\epsilon \in (0, 1)$ , let  $\Pi = SD$  where  $S \in \mathbb{R}^{s \times n}$  has each column chosen independently and uniformly from the  $s$  standard basis vectors of  $\mathbb{R}^s$  and  $D \in \mathbb{R}^{n \times n}$  is a diagonal matrix with diagonal entries chosen independently and uniformly from  $\pm 1$ . Let  $s = (d^4 + 3d^3)/\epsilon^2$ . Then with probability at least 0.5,*

$$(1 - \epsilon)\|Ax\|_2 \leq \|\Pi Ax\|_2 \leq (1 + \epsilon)\|Ax\|_2, \quad \forall x \in \mathbb{R}^d.$$

*In addition,  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A))$  time.*

The construction of  $\Pi$  in this theorem is the same as the construction in Clarkson and Woodruff [11]. For them,  $s = \mathcal{O}((d/\epsilon)^4 \log^2(d/\epsilon))$  in order to achieve  $(1 \pm \epsilon)$  distortion with a constant probability. Theorem 1 shows that it actually suffices to set  $s = (d^4 + 3d^3)/\epsilon^2$ , thus nailing down the constant factors in the big- $\mathcal{O}$ , removing the  $\log^2(d/\epsilon)$  factor, and improving the dependence on  $\epsilon$  from  $1/\epsilon^4$  to  $1/\epsilon^2$ . Surprisingly, the proof is rather simple: we apply Chebyshev's inequality to the elements of  $X = U^T \Pi^T \Pi U$ , where  $U$  is an orthonormal basis for  $\mathcal{A}_2$ , to obtain a union bound on  $\|X - I\|_\infty$ , and we then use Gershgorin's circle theorem to bound the eigenvalues of  $X$  such that  $\|X - I\|_2 \leq \epsilon$ . See Appendix A.1 for a complete proof.

**Remark.** The  $\mathcal{O}(\text{nnz}(A))$  running time is indeed optimal, up to constant factors, for general inputs. Consider the case when  $A$  has an important row  $a_j$  such that  $A$  becomes rank-deficient without it. Thus, we have to observe  $a_j$  in order to compute a low-distortion embedding. However, without any prior knowledge, we have to scan at least a constant portion of the input to guarantee that  $a_j$  is observed with a constant probability, which takes  $\mathcal{O}(\text{nnz}(A))$  time. Note that this optimality result applies to general  $p$ .

The results of Theorem 1 propagate to related applications, e.g., to the  $\ell_2$  regression problem, the low-rank matrix approximation problem and the problem of computing approximations to the  $\ell_2$  leverage scores. Since it underlies the other applications, only the  $\ell_2$  regression improvement is stated here explicitly; its proof is basically combining our Theorem 1 with Theorem 19 of [11].

**Corollary 1** (Fast  $\ell_2$  Regression). *With a constant probability, a  $(1 + \epsilon)$ -approximate solution to an  $\ell_2$  regression problem can be computed in  $\mathcal{O}(\text{nnz}(A) + \mathcal{T}_2(\epsilon; d^4/\epsilon^2, d))$  time.*

**Remark.** Although our simpler direct proof leads to a better result for  $\ell_2$  subspace embedding, the technique used in the proof of Clarkson and Woodruff [11], which splits coordinates into “heavy” and “light” sets based on the leverage scores, highlights an important structural property of  $\ell_2$  subspace: that only a small subset of coordinates can have large  $\ell_2$  leverage scores. (We note that the technique of splitting coordinates is also used by Ailon and Liberty [1] to get an unrestricted fast Johnson-Lindenstrauss transform; and that the difficulty in finding and approximating the large-leverage directions was—until recently [20, 15]—responsible for difficulties in obtaining fast relative-error random sampling algorithms for  $\ell_2$  regression and low-rank matrix approximation.) An analogous structural fact holds for  $\ell_1$  and other  $\ell_p$  spaces. Using this property, we can construct novel input-sparsity time  $\ell_p$  subspace embeddings for general  $p \in [1, 2)$ , as we discuss in the next two sections.

### 4 Main Results for $\ell_1$ Embedding

Here is our main result for input-sparsity time low-distortion subspace embeddings for  $\ell_1$ .

**Theorem 2** (Low-distortion Embedding for  $\ell_1$ ). *Given  $A \in \mathbb{R}^{n \times d}$  with full column rank, let  $\Pi = SC \in \mathbb{R}^{s \times n}$ , where  $S \in \mathbb{R}^{s \times n}$  has each column chosen independently and uniformly from the  $s$  standard basis vectors of  $\mathbb{R}^s$ , and where  $C \in \mathbb{R}^{n \times n}$  is a diagonal matrix with diagonals chosen independently from the standard Cauchy distribution. Set  $s = \omega d^5 \log^5 d$  with  $\omega$  sufficiently large. Then with a constant probability, we have*

$$1/\mathcal{O}(d^2 \log^2 d) \cdot \|Ax\|_1 \leq \|\Pi Ax\|_1 \leq \mathcal{O}(d \log d) \cdot \|Ax\|_1, \quad \forall x \in \mathbb{R}^d.$$

*In addition,  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A))$  time.*

The construction of the  $\ell_1$  subspace embedding matrix is different than its  $\ell_2$  norm counterpart only by the diagonal elements of  $D$  (or  $C$ ): whereas we use  $\pm 1$  for the  $\ell_2$  norm, we use Cauchy variables for the  $\ell_1$  norm. The proof of Theorem 2 uses the technique of splitting coordinates, the fact that the Cauchy distribution is 1-stable, and the upper and lower tail inequalities regarding the Cauchy distribution from Lemmas 5 and 6. See Appendix A.2 for a complete proof.

**Remark.** As mentioned above, the  $\mathcal{O}(\text{nnz}(A))$  running time is optimal. Whether the distortion  $\mathcal{O}(d^3 \log^3 d)$  is optimal is still an open question. However, for the same construction of  $\Pi$ , we can provide a “bad” case that provides a lower bound. Choose  $A = (I_d \quad \mathbf{0})^T$ . Suppose that  $s$  is sufficiently large such that with an overwhelming probability, the top  $d$  rows of  $A$  are perfectly hashed, i.e.,  $\|\Pi Ax\|_1 = \sum_{k=1}^d |c_k| |x_k|$ ,  $\forall x \in \mathbb{R}^d$ , where  $c_k$  is the  $k$ -th diagonal of  $C$ . Then, the distortion of  $\Pi$  is  $\max_{k \leq d} |c_k| / \min_{k \leq d} |c_k| \approx \mathcal{O}(d^2)$ . Therefore, at most an  $\mathcal{O}(d \log^3 d)$  factor of the distortion is due to artifacts in our analysis.

Our input-sparsity time  $\ell_1$  subspace embedding of Theorem 2 improves the  $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$ -time embedding by Sohler and Woodruff [28] and the  $\mathcal{O}(nd \log n)$ -time embedding of Clarkson et al. [10]. In addition, by combining Theorem 2 and Lemma 3, we can compute a  $(1 \pm \epsilon)$ -distortion embedding in  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time, i.e., in *nearly* input-sparsity time.

**Theorem 3**  $((1 \pm \epsilon)$ -distortion Embedding for  $\ell_1$ ). *Given  $A \in \mathbb{R}^{n \times d}$ , it takes  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time to compute a sampling matrix  $S \in \mathbb{R}^{s \times n}$  with  $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$  such that with a constant probability,  $S$  embeds  $\mathcal{A}_1$  into  $(\mathbb{R}^s, \|\cdot\|_1)$  with distortion  $1 \pm \epsilon$ .*

Our improvements in Theorems 2 and 3 also propagate to related  $\ell_1$ -based applications, including the  $\ell_1$  regression and the  $\ell_1$  subspace approximation problem considered in [28, 10]. As before, only the regression improvement is stated here explicitly. For completeness, we present in Algorithm 1 our algorithm for solving  $\ell_1$  regression problems in nearly input-sparsity time. The brief proof of Corollary 2, our main quality-of-approximation result for Algorithm 1, may be found in Appendix A.3.

---

**Algorithm 1** Fast  $\ell_1$  Regression Approximation in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) \log(1/\epsilon)/\epsilon^2)$  Time

---

**Input:**  $A \in \mathbb{R}^{n \times d}$  with full column rank,  $b \in \mathbb{R}^n$ , and  $\epsilon \in (0, 1/2)$ .

**Output:** A  $(1 + \epsilon)$ -approximation solution  $\hat{x}$  to  $\min_{x \in \mathbb{R}^d} \|Ax - b\|_1$ , with a constant probability.

- 1: Let  $\bar{A} = (A \quad b)$  and denote  $\bar{\mathcal{A}}_1$  the  $\ell_1$  subspace spanned by  $A$ ’s columns and  $b$ .
  - 2: Compute a low-distortion embedding  $\Pi \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$  of  $\bar{\mathcal{A}}_1$  (Theorem 2).
  - 3: Compute  $\bar{R} \in \mathbb{R}^{(d+1) \times (d+1)}$  from  $\Pi \bar{A}$  such that  $\bar{A} \bar{R}^{-1}$  is well-conditioned (QR or Lemma 2).
  - 4: Compute a  $(1 \pm \epsilon/4)$ -distortion embedding  $S \in \mathbb{R}^{\mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2) \times n}$  of  $\bar{\mathcal{A}}_1$  (Lemma 3).
  - 5: Compute a  $(1 + \epsilon/4)$ -approximate solution  $\hat{x}$  to  $\min_{x \in \mathbb{R}^d} \|SAx - Sb\|_1$ .
-



**Corollary 2** (Fast  $\ell_1$  Regression). *With a constant probability, Algorithm 1 computes a  $(1 + \epsilon)$ -approximate solution to an  $\ell_1$  regression problem in  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_1(\epsilon; \text{poly}(d) \log(1/\epsilon)/\epsilon^2, d))$  time.*

**Remark.** For readers familiar with the impossibility results for dimension reduction in  $\ell_1$  [8, 18, 5], note that those results apply to arbitrary point sets of size  $n$  and are interested in embeddings that are “oblivious,” in that they do not depend on the input data. In this paper, we only consider points in a subspace, and the subspace-preserving sampling procedure of [12] that we use is data-dependent.

## 5 Main Results for $\ell_p$ Embedding

In this section, we use the properties of  $p$ -stable distributions to generalize the input-sparsity time  $\ell_1$  subspace embedding to  $\ell_p$  norms, for  $p \in (1, 2)$ . Generally,  $\mathcal{D}_p$  does not have explicit PDF/CDF, which increases the difficulty for theoretical analysis. Indeed, the main technical difficulty here is that we are not aware of  $\ell_p$  analogues of Lemmas 5 and 6 that would provide upper and lower tail inequality for  $p$ -stable distributions. (Indeed, even Lemmas 5 and 6 were established only recently [10].)

Instead of analyzing  $\mathcal{D}_p$  directly, for any  $p \in (1, 2)$ , we establish an order among the Cauchy distribution, the  $p$ -stable distribution, and the Gaussian distribution, and then we derive upper and lower tail inequalities for the  $p$ -stable distribution similar to the ones we used to prove Theorem 2. We state these technical results here since they are of independent interest. We start with the following lemma, which is proved in Appendix A.4 and which establishes this order.

**Lemma 8.** *For any  $p \in (1, 2)$ , there exist constants  $\alpha_p > 0$  and  $\beta_p > 0$  such that*

$$\alpha_p |C| \succeq |X_p|^p \succeq \beta_p |G|^2,$$

where  $C$  is a standard Cauchy variable,  $X_p \sim \mathcal{D}_p$ ,  $G$  is a standard Gaussian variable. By “ $X \succeq Y$ ” we mean  $\Pr[X \geq t] \geq \Pr[Y \geq t]$ ,  $\forall t \in \mathbb{R}$ , i.e.,  $F_X(t) \leq F_Y(t)$ ,  $\forall t \in \mathbb{R}$ , where  $F(\cdot)$  is the corresponding CDF.

Our numerical results suggest that the constants  $\alpha_p$  and  $\beta_p$  are not too far away from 1. See Figure 1, which plots of the CDFs of  $|X_p/2|^p$  for  $p = 1, 0, 1.1, \dots, 2.0$ , based on which we conjecture  $|X_{p_1}/2|^{p_1} \succeq |X_{p_2}/2|^{p_2}$ , for all  $1 \leq p_1 \leq p_2 \leq 2$ . This implies that  $2^{p-1}|C| \succeq |X_p|^p$  and  $|X_p|^p \succeq 2^{p-2}|X_2|^2 \simeq 2^{p-1}|G|^2$ , which therefore provides a value for the constants  $\alpha_p$  and  $\beta_p$ .

Lemma 8 suggests that we can use Lemma 5 (regarding Cauchy random variables) to derive upper tail inequalities for general  $p$ -stable distributions and that we can use Lemma 7 (regarding Gaussian variables) to derive lower tail inequalities for general  $p$ -stable distributions. The following two lemmas establish these results; the proofs of these lemmas are provided in Appendix A.5 and Appendix A.6, respectively.

**Lemma 9** (Upper Tail Inequality for  $p$ -stable Distributions). *Given  $p \in (1, 2)$ , for  $i = 1, \dots, m$ , let  $X_i$  be  $m$  (not necessarily independent) random variables sampled from  $\mathcal{D}_p$ , and  $\gamma_i > 0$  with  $\gamma = \sum_i \gamma_i$ . Let  $X = \sum_i \gamma_i |X_i|^p$ . Assume that  $m \geq 3$ . Then for any  $t \geq 1$ ,*

$$\Pr[X \geq t\alpha_p\gamma] \leq \frac{2\log(mt)}{t}.$$

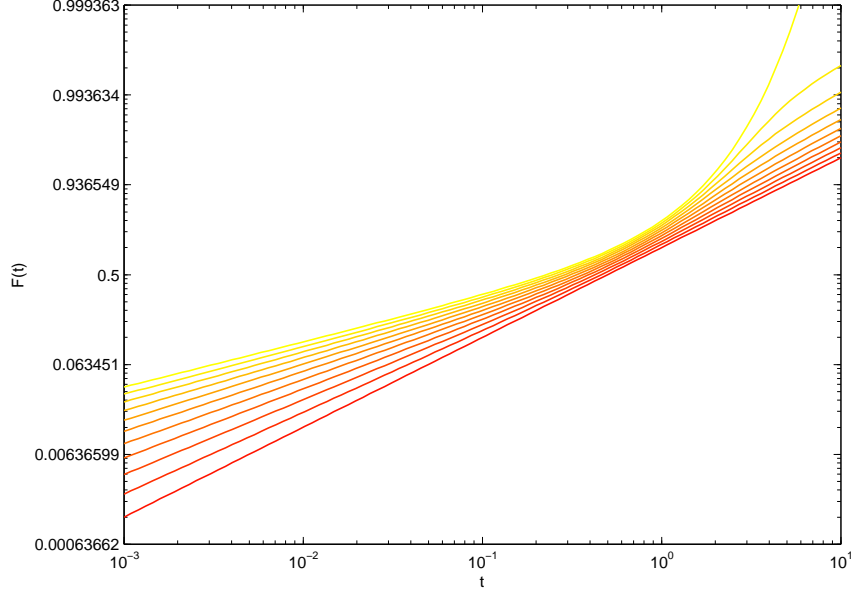


Figure 1: The CDFs ( $F(t)$ ) of  $|X_p/2|^p$  for  $p = 1.0$  (bottom, i.e., red or dark gray),  $1.1, \dots, 2.0$  (top, i.e., yellow or light gray), where  $X_p \sim \mathcal{D}_p$  and the scales of the axes are chosen to magnify the upper (as  $t \rightarrow \infty$ ) and lower (as  $t \rightarrow 0$ ) tails. These empirical results suggest  $|X_{p_1}/2|^{p_1} \succeq |X_{p_2}/2|^{p_2}$  for all  $1 \leq p_1 \leq p_2 \leq 2$ .

**Lemma 10** (Lower Tail Inequality for  $p$ -stable Distributions). *For  $i = 1, \dots, m$ , let  $X_i$  be independent random variables sampled from  $\mathcal{D}_p$ , and  $\gamma_i \geq 0$  with  $\gamma = \sum_i \gamma_i$ . Let  $X = \sum_i \gamma_i |c_i|$ . Then,*

$$\log \Pr[X \leq (1-t)\beta_p \gamma] \leq \frac{-\gamma t^2}{6 \max_i \gamma_i}.$$

Given these results, here is our main result for input-sparsity time low-distortion subspace embeddings for  $\ell_p$ . The proof of this theorem is similar to the proof of Theorem 2, except that we replace the  $\ell_1$  norm  $\|\cdot\|_1$  by  $\|\cdot\|_p^p$  and use the tail inequalities from Lemmas 9 and 10 (rather than Lemmas 5 and 6).

**Theorem 4** (Low-distortion Embedding for  $\ell_p$ ). *Given  $A \in \mathbb{R}^{n \times d}$  with full column rank and  $p \in (1, 2)$ , let  $\Pi = SD \in \mathbb{R}^{s \times n}$  where  $S \in \mathbb{R}^{s \times n}$  has each column chosen independently and uniformly from the  $s$  standard basis vectors of  $\mathbb{R}^s$ , and where  $D \in \mathbb{R}^{n \times n}$  is a diagonal matrix with diagonals chosen independently from  $\mathcal{D}_p$ . Set  $s = \omega d^5 \log^5 d$  with  $\omega$  sufficiently large. Then with a constant probability, we have*

$$1/\mathcal{O}((d \log d)^{2/p}) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}((d \log d)^{1/p}) \cdot \|Ax\|_p, \quad \forall x \in \mathbb{R}^d.$$

*In addition,  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A))$  time.*

Similar to the  $\ell_1$  case, our input-sparsity time  $\ell_p$  subspace embedding of Theorem 4 improves the  $\mathcal{O}(nd \log n)$ -time embedding of Clarkson et al. [10]. In addition, by combining Theorem 4 and Lemma 3, we can compute a  $(1 \pm \epsilon)$ -distortion embedding in  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time.

**Theorem 5**  $((1 \pm \epsilon)$ -distortion Embedding for  $\ell_p$ ). *Given  $A \in \mathbb{R}^{n \times d}$  and  $p \in [1, 2)$ , it takes  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time to compute a sampling matrix  $S \in \mathbb{R}^{s \times n}$  with  $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$  such that with a constant probability,  $S$  embeds  $\mathcal{A}_p$  into  $(\mathbb{R}^s, \|\cdot\|_p)$  with distortion  $1 \pm \epsilon$ .*

These improvements for  $\ell_p$  subspace embedding also propagate to related  $\ell_p$ -based applications. In particular, we can establish an improved algorithm for solving the  $\ell_p$  regression problem in nearly input-sparsity time.

**Corollary 3** (Fast  $\ell_p$  Regression). *Given  $p \in (1, 2)$ , with a constant probability, a  $(1 + \epsilon)$ -approximate solution to an  $\ell_p$  regression problem can be computed in*

$$\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_p(\epsilon; \text{poly}(d) \log(1/\epsilon)/\epsilon^2, d))$$

*time.*

For completeness, we also present a result for low-distortion dense embeddings for  $\ell_p$  that the tail inequalities from Lemmas 9 and 10 enable us to construct. See Appendix A.7 for a proof of the following theorem.

**Theorem 6** (Low-distortion Dense Embedding for  $\ell_p$ ). *Given  $A \in \mathbb{R}^{n \times d}$  with full column rank and  $p \in (1, 2)$ , let  $\Pi \in \mathbb{R}^{s \times n}$  whose entries are i.i.d. samples from  $\mathcal{D}_p$ . If  $s = \omega d \log d$  for  $\omega$  sufficiently large, with a constant probability, we have*

$$1/\mathcal{O}(1) \cdot \|Ax\|_p \leq \|\Pi Ax\|_p \leq \mathcal{O}((d \log d)^{1/p}) \cdot \|Ax\|_p, \quad \forall x \in \mathbb{R}^d.$$

*In addition,  $\Pi A$  can be computed in  $\mathcal{O}(\text{nnz}(A) \cdot d \log d)$  time.*

**Remark.** The result in Theorem 6 is based on a dense  $\ell_p$  subspace embeddings that is analogous to the dense Gaussian embedding for  $\ell_2$  and the dense Cauchy embedding of [28] for  $\ell_1$ . Although the running time (if one is simply interested in FLOP counts in RAM) of Theorem 6 is somewhat worse than that of Theorem 4, the embedding dimension and condition number quality (the ratio of the upper bound on the distortion and the lower bound on the distortion) are much better. Our numerical implementations, both with the  $\ell_1$  norm [10] and with the  $\ell_2$  norm [23], strongly suggest that the latter quantities are more important to control when implementing randomized regression algorithms in large-scale parallel and distributed settings.

## 6 Improving the Embedding Dimension

In Theorem 2 and Theorem 4, the embedding dimension is  $s = \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2)$ , where the  $\text{poly}(d)$  term is a somewhat large polynomial of  $d$  that directly multiplies the  $\log(1/\epsilon)/\epsilon^2$  term. (See the remark below for comments on the precise value of the  $\text{poly}(d)$  term.) This is not ideal for the subspace embedding and the  $\ell_p$  regression, because we want to have a small embedding dimension and a small subsampled problem, respectively. Here, we show that it is possible to decouple the large polynomial of  $d$  and the  $\log(1/\epsilon)/\epsilon^2$  term via another round of sampling and conditioning without increasing the complexity. See Algorithm 2 for details on this procedure. Theorem 7 provides our main quality-of-approximation result for Algorithm 2; its proof can be found in Appendix A.8.

**Theorem 7** (Improving the Embedding Dimension). *Given  $p \in [1, 2)$ , with a constant probability, Algorithm 2 computes a  $(1 \pm \epsilon)$ -distortion embedding of  $\mathcal{A}_p$  into  $(\mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2)}, \|\cdot\|_p)$  in  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time.*

Then, by applying Theorem 7 to the  $\ell_p$  regression problem, we can improve the size of the subsampled problem and hence the overall running time.

---

**Algorithm 2** Improving the Embedding Dimension

---

**Input:**  $A \in \mathbb{R}^{n \times d}$  with full column rank,  $p \in [1, 2)$ , and  $\epsilon \in (0, 1)$ .

**Output:** A  $(1 \pm \epsilon)$ -distortion embedding  $S \in \mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times n}$  of  $\mathcal{A}_p$ .

- 1: Compute a low-distortion embedding  $\tilde{\Pi} \in \mathbb{R}^{\mathcal{O}(\text{poly}(d)) \times n}$  of  $\mathcal{A}_p$  (Theorems 2 and 4).
  - 2: Compute  $\tilde{R} \in \mathbb{R}^{d \times d}$  from  $\tilde{\Pi}A$  such that  $A\tilde{R}^{-1}$  is well-conditioned (QR or Lemma 2).
  - 3: Compute a  $(1 \pm 1/2)$ -distortion embedding  $\tilde{S} \in \mathbb{R}^{\mathcal{O}(\text{poly}(d) \times n)}$  of  $\mathcal{A}_p$  (Lemma 3).
  - 4: Compute  $R \in \mathbb{R}^{d \times d}$  such that  $\kappa_p(\tilde{S}AR^{-1}) \leq 2d$  (Theorem 2).
  - 5: Compute a  $(1 \pm \epsilon)$ -distortion embedding  $S \in \mathbb{R}^{\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times n}$  of  $\mathcal{A}_p$  (Lemma 3).
- 

**Corollary 4** (Improved Fast  $\ell_p$  Regression). *Given  $p \in [1, 2)$ , with a constant probability, a  $(1 + \epsilon)$ -approximate solution to an  $\ell_p$  regression problem can be computed in*

$$\mathcal{O}(\text{nnz}(A) \cdot \log n + \mathcal{T}_p(\epsilon; d^{3+p/2} \log(1/\epsilon)/\epsilon^2, d))$$

*time. The second term comes from solving a subsampled problem of size  $\mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2) \times d$ .*

**Remark.** We have stated our results in the previous sections as  $\text{poly}(d)$  without stating the value of the polynomial because there are numerous trade-offs between the conditioning quality and the running time. For example, let  $p = 1$ . We can use a rounding algorithm instead of QR to compute the  $R$  matrix. If we use the input-sparsity time embedding with the  $\mathcal{O}(d)$ -rounding algorithm of [10], then the running time to compute the  $(1 \pm \epsilon)$ -distortion embedding is  $\mathcal{O}(\text{nnz}(A) \cdot \log n + d^8/\epsilon^2)$  and the embedding dimension is  $\mathcal{O}(d^{6.5}/\epsilon^2)$  (ignoring  $\log$  factors). If, on the other hand, we use QR to compute  $R$ , then the running time is  $\mathcal{O}(\text{nnz}(A) \cdot \log n + d^7/\epsilon^2)$  and the embedding dimension is  $\mathcal{O}(d^8/\epsilon^2)$ . However, with the result from this section, the running time is simply  $\mathcal{O}(\text{nnz}(A) \cdot \log n + \text{poly}(d) + \mathcal{T}_p(\epsilon; d^{3+p/2}/\epsilon^2, d))$  and the  $\text{poly}(d)$  term can be absorbed by the  $\text{nnz}(A)$  term.

## References

- [1] N. Ailon and E. Liberty. An almost optimal unrestricted fast Johnson-Lindenstrauss transform. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 185–191, 2011.
- [2] H. Auerbach. *On the area of convex curves with conjugate diameters*. PhD thesis, University of Lwów, 1930.
- [3] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging LAPACK’s least-squares solver. *SIAM Journal on Scientific Computing*, 32:1217–1236, 2010.
- [4] J. Bourgain, J. Lindenstrauss, and V. Milman. Approximation of zonoids by zonotopes. *Acta Mathematica*, 162:73–141, 1989.
- [5] B. Brinkman and M. Charikar. On the impossibility of dimension reduction in  $\ell_1$ . *Journal of the ACM*, 52(5):766–788, 2005.
- [6] J. P. Brooks and J. H. Dulá. The L1-norm best-fit hyperplane problem. *Applied Mathematics Letters*, 26(1):51–55, 2013.
- [7] J. M. Chambers, C. L. Mallows, and B. W. Stuck. A method for simulating stable random variables. *Journal of the American Statistical Association*, 71(354):340–344, 1976.

- [8] M. Charikar and A. Sahai. Dimension reduction in the  $\ell_1$  norm. In *Proceedings of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, pages 551–560, 2002.
- [9] K. Clarkson. Subgradient and sampling algorithms for  $\ell_1$  regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 257–266, 2005.
- [10] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The Fast Cauchy Transform and faster robust linear regression. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, 2013.
- [11] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. Technical report. Preprint: arXiv:1207.6365 (2012).
- [12] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. Sampling algorithms and coresets for  $\ell_p$  regression. *SIAM Journal on Computing*, (38):2060–2078, 2009.
- [13] A. Dasgupta, R. Kumar, and T. Sarlós. A sparse Johnson-Lindenstrauss transform. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing*, pages 341–350, 2010.
- [14] K. R. Davidson and S. J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pages 317–366. North Holland, 2001.
- [15] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [16] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for  $\ell_2$  regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1127–1136, 2006.
- [17] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, 117(2):219–249, 2010.
- [18] J. R. Lee and A. Naor. Embedding the diamond graph in  $L_p$  and dimension reduction in  $L_1$ . *Geometric And Functional Analysis*, 14(4):745–747, 2004.
- [19] P. Lévy. *Calcul des Probabilités*. Gauthier-Villars, Paris, 1925.
- [20] M. W. Mahoney. *Randomized algorithms for matrices and data*. Foundations and Trends in Machine Learning. NOW Publishers, Boston, 2011. Also available at: arXiv:1104.5557.
- [21] M. W. Mahoney and P. Drineas. CUR matrix decompositions for improved data analysis. *Proc. Natl. Acad. Sci. USA*, 106:697–702, 2009.
- [22] A. Maurer. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):Article 15, 2003.
- [23] X. Meng, M. A. Saunders, and M. W. Mahoney. LSRN: A parallel iterative solver for strongly over- or under-determined systems. Technical report. Preprint: arXiv:1109.5981 (2011).
- [24] J. E. Mitchell. Polynomial interior point cutting plane methods. *Optimization Methods and Software*, 18(5):507–534, 2003.

- [25] J. P. Nolan. *Stable Distributions - Models for Heavy Tailed Data*. Birkhauser, Boston, 2012. In progress, Chapter 1 online at [academic2.american.edu/~jpnolan](http://academic2.american.edu/~jpnolan).
- [26] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proc. Natl. Acad. Sci. USA*, 105(36):13212–13217, 2008.
- [27] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, 2006.
- [28] C. Sohler and D. P. Woodruff. Subspace embeddings for the  $\ell_1$ -norm with applications. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pages 755–764, 2011.

## A Appendix

### A.1 Proof of Theorem 1 ((1 ± ε)-distortion Embedding for $\ell_2$ )

Let the  $n \times d$  matrix  $U$  be an orthonormal basis for the range of the  $n \times d$  matrix  $A$ . Rather than proving the theorem by establishing that

$$(1 - \epsilon)\|Uz\|_2 \leq \|\Pi Uz\|_2 \leq (1 + \epsilon)\|Uz\|_2$$

holds for all  $z \in \mathbb{R}^d$ , as is essentially done in, e.g., [16] and [11], we note that  $U^T U = I_d$ , and we directly bound the extent to which the embedding process perturbs this product. To do so, define

$$X = (\Pi U)^T (\Pi U) = U^T D^T S^T S D U.$$

That is,

$$x_{kl} = \sum_{i=1}^s \left( \sum_{j=1}^n s_{ij} d_j u_{jk} \right) \left( \sum_{j=1}^n s_{ij} d_j u_{jl} \right), \quad k, l \in \{1, \dots, d\},$$

where  $s_{ij}$  is the  $(i, j)$ -th element of  $S$ ,  $d_j$  is the  $j$ -th diagonal element of  $D$ , and  $u_{jk}$  is the  $(j, k)$ -th element of  $U$ . We will use the following facts in the proof:

$$\begin{aligned} \mathbf{E}[d_{j_1} d_{j_2}] &= \delta_{j_1 j_2}, \\ \mathbf{E}[s_{i_1 j_1} s_{i_2 j_2}] &= \begin{cases} \frac{1}{s^2} & \text{if } j_1 \neq j_2, \\ \frac{1}{s} & \text{if } i_1 = i_2, j_1 = j_2, \\ 0 & \text{if } i_1 \neq i_2, j_1 = j_2. \end{cases} \end{aligned}$$

We have,

$$\mathbf{E}[x_{kl}] = \sum_i \sum_{j_1, j_2} \mathbf{E}[s_{i j_1} d_{j_1} u_{j_1 k} \cdot s_{i j_2} d_{j_2} u_{j_2 l}] = \sum_i \sum_j \mathbf{E}[s_{i j} u_{j k} u_{j l}] = \sum_j u_{j k} u_{j l} = \delta_{kl},$$

and we also have

$$\mathbf{E}[x_{kl}^2] = \mathbf{E} \left[ \left( \sum_i \left( \sum_j s_{i j} d_j u_{j k} \right) \left( \sum_j s_{i j} d_j u_{j l} \right) \right)^2 \right]$$

$$\begin{aligned}
&= \sum_{i_1, i_2} \mathbf{E} \left[ \left( \sum_j s_{i_1 j} d_j u_{jk} \right) \left( \sum_j s_{i_1 j} d_j u_{jl} \right) \left( \sum_j s_{i_2 j} d_j u_{jk} \right) \left( \sum_j s_{i_2 j} d_j u_{jl} \right) \right] \\
&= \sum_{i_1, i_2} \sum_{j_1, j_2, j_3, j_4} \mathbf{E}[s_{i_1 j_1} d_{j_1} u_{j_1 k} \cdot s_{i_1 j_2} d_{j_2} u_{j_2 l} \cdot s_{i_2 j_3} d_{j_3} u_{j_3 k} \cdot s_{i_2 j_4} d_{j_4} u_{j_4 l}] \\
&= \sum_{i_1, i_2} \left( \sum_j \mathbf{E}[s_{i_1 j} u_{jk} \cdot s_{i_1 j} u_{jl} \cdot s_{i_2 j} u_{jk} \cdot s_{i_2 j} u_{jl}] \right. \\
&\quad + \sum_{j_1 \neq j_2} \mathbf{E}[s_{i_1 j_1} u_{j_1 k} \cdot s_{i_1 j_1} u_{j_1 l} \cdot s_{i_2 j_2} u_{j_2 k} \cdot s_{i_2 j_2} u_{j_2 l}] \\
&\quad + \sum_{j_1 \neq j_2} \mathbf{E}[s_{i_1 j_1} u_{j_1 k} \cdot s_{i_1 j_2} u_{j_2 l} \cdot s_{i_2 j_1} u_{j_1 k} \cdot s_{i_2 j_2} u_{j_2 l}] \\
&\quad \left. + \sum_{j_1 \neq j_2} \mathbf{E}[s_{i_1 j_1} u_{j_1 k} \cdot s_{i_1 j_2} u_{j_2 l} \cdot s_{i_2 j_2} u_{j_2 k} \cdot s_{i_2 j_1} u_{j_1 l}] \right) \\
&= \sum_j u_{jk}^2 u_{jl}^2 + \sum_{j_1 \neq j_2} u_{j_1 k} u_{j_1 l} u_{j_2 k} u_{j_2 l} + \frac{1}{s} \sum_{j_1 \neq j_2} u_{j_1 k}^2 u_{j_2 l}^2 + \frac{1}{s} \sum_{j_1 \neq j_2} u_{j_1 k} u_{j_2 l} u_{j_2 k} u_{j_1 l} \\
&= \left( \sum_j u_{jk} u_{jl} \right)^2 + \frac{1}{s} \left( \left( \sum_j u_{jk}^2 \right) \left( \sum_j u_{jl}^2 \right) + \left( \sum_j u_{jk} u_{jl} \right)^2 - 2 \sum_j u_{jk}^2 u_{jl}^2 \right) \\
&= \begin{cases} 1 + \frac{2}{s} (1 - \|U_{*k}\|_4^4) & \text{if } k = l, \\ \frac{1}{s} (1 - 2\langle U_{*k}^2, U_{*l}^2 \rangle) & \text{if } k \neq l. \end{cases}
\end{aligned}$$

By Chebyshev's inequality, it follows that

$$\mathbf{Pr}[|x_{kl} - \delta_{kl}| \geq t] \leq \frac{\mathbf{var}[x_{kl}]}{t^2} = \frac{\mathbf{E}[x_{kl}^2] - \delta_{kl}^2}{t^2} \leq \begin{cases} \frac{2}{st^2} & \text{if } k = l, \\ \frac{1}{st^2} & \text{if } k \neq l. \end{cases}$$

Denote by  $\mathcal{E}$  the event  $|X - I|_\infty \leq \frac{\epsilon}{d}$ , where recall that  $|\cdot|_\infty$  denotes the element-wise  $\ell_\infty$  norm of a matrix. Note that  $X$  is symmetric. We have,

$$\begin{aligned}
\mathbf{Pr}[\mathcal{E}] &\geq 1 - \sum_{k \leq l} \mathbf{Pr}\left[|x_{kl} - \delta_{kl}| \geq \frac{\epsilon}{d}\right] \\
&= 1 - \sum_k \frac{2d^2}{s\epsilon^2} - \sum_{k < l} \frac{d^2}{s\epsilon^2} \\
&= 1 - \frac{2d^3 + \frac{d^3(d-1)}{2}}{s\epsilon^2}.
\end{aligned}$$

If we set  $s = (d^4 + 3d^3)/\epsilon^2$ , then we have  $\mathbf{Pr}[\mathcal{E}] \geq 0.5$ . By Gershgorin's circle theorem,  $\mathcal{E}$  implies  $1 - \epsilon \leq \lambda(X) \leq 1 + \epsilon$ , i.e.,

$$(1 - \epsilon)\|z\|_2^2 \leq z^T X z \leq (1 + \epsilon)\|z\|_2^2, \quad \forall z \in \mathbb{R}^d,$$

which is equivalent to

$$\sqrt{1 - \epsilon}\|z\|_2 \leq \|\Pi U z\|_2 \leq \sqrt{1 + \epsilon}\|z\|_2, \quad \forall z \in \mathbb{R}^d,$$

which, since  $1 - \epsilon < \sqrt{1 - \epsilon}$  and  $1 + \epsilon > \sqrt{1 + \epsilon}$  for any  $\epsilon \in (0, 1)$ , implies that

$$(1 - \epsilon)\|z\|_2 \leq \|\Pi U z\|_2 \leq (1 + \epsilon)\|z\|_2, \quad \forall z \in \mathbb{R}^d.$$

Therefore,  $\mathcal{E}$  implies

$$(1 - \epsilon)\|Ax\|_2 \leq \|\Pi Ax\|_2 \leq (1 + \epsilon)\|Ax\|_2, \quad \forall x \in \mathbb{R}^d,$$

which concludes the proof.

## A.2 Proof of Theorem 2 (Low-distortion Embedding for $\ell_1$ )

We start with the following result, which establishes the existence of the so-called Auerbach's basis of a  $d$ -dimensional normed vector space. For our proof, we will only need its existence and not an algorithm to construct it.

**Lemma 11.** (*Auerbach [2]*) *Let  $(\mathcal{A}, \|\cdot\|)$  be a  $d$ -dimensional normed vector space. There exists a basis  $\{e_1, \dots, e_d\}$  of  $\mathcal{A}$ , called Auerbach basis, such that  $\|e_k\| = 1$  and  $\|e^k\|^* = 1$  for  $k = 1, \dots, d$ , where  $\{e^1, \dots, e^n\}$  is a basis of  $\mathcal{A}^*$  dual to  $\{e_1, \dots, e_n\}$ .*

This Auerbach's lemma implies that a  $(d, 1, 1)$ -conditioned basis matrix of  $\mathcal{A}_1$  exists, which will be denoted by  $U$  throughout the proof. By definition,  $U$ 's columns are unit vectors in the  $\ell_1$  norm (thus  $|U|_1 = d$ , where recall that  $|\cdot|_1$  denotes the element-wise  $\ell_1$  norm of a matrix) and  $\|x\|_\infty \leq \|Ux\|_1$ ,  $\forall x \in \mathbb{R}^d$ . Denote by  $u_j$  the  $j$ -th row of  $U$ ,  $j = 1, \dots, n$ . Define  $v_j = \|u_j\|_1$  the  $\ell_1$  leverage scores of  $A$ . We have  $\sum_j v_j = |U|_1 = d$ . Let  $\tau > 0$  to be determined later, and define two index sets  $H = \{j \mid v_j \geq \tau\}$  and  $L = \{j \mid v_j < \tau\}$ . It is easy to see that  $|H| \leq \frac{d}{\tau}$  where  $|\cdot|$  is used to denote the size of a finite set, and  $\|v^L\|_\infty \leq \tau$  where

$$v_j^L = \begin{cases} v_j, & \text{if } j \in L \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, \dots, n.$$

Similarly, when an index set appears as a superscript, we mean zeroing out elements or rows that do not belong to this index set, e.g.,  $v^L$  and  $U^L$ . Define

$$Y = \{y \in \mathbb{R}^n \mid y = Ux, \|x\|_\infty = 1, x \in \mathbb{R}^d\}.$$

For any  $y = Ux \in Y$ , we have  $\|y\|_1 = \|Ux\|_1 \geq \|x\|_\infty = 1$ ,

$$|y_j| = |u_j^T x| \leq \|u_j\|_1 \|x\|_\infty = v_j, \quad j = 1, \dots, n,$$

and thus  $\|y\|_1 \leq \|v\|_1 = d$ . Define  $Y^L = \{y \in Y \mid \|y^L\|_1 \geq \frac{1}{2}\|y\|_1\}$  and  $Y^H = Y \setminus Y^L$ . Given  $S$ , define a mapping  $\phi : \{1, \dots, n\} \rightarrow \{1, \dots, s\}$  such that  $s_{\phi(j), j} = 1$ ,  $j = 1, \dots, n$ , and split  $L$  into two subsets:  $\hat{L} = \{j \in L \mid \phi(j) \in \phi(H)\}$  and  $\bar{L} = L \setminus \hat{L}$ . Consider these events:

- $\mathcal{E}_U$ :  $|\Pi U|_1 \leq \omega_1 d \log d$  for some  $\omega_1 > 0$ .
- $\mathcal{E}_L$ :  $\|Sv^L\|_\infty \leq \omega_2/(d \log d)$  for some  $\omega_2 > 0$ .
- $\mathcal{E}_H$ :  $\phi(j_1) \neq \phi(j_2)$ ,  $\forall j_1 \neq j_2$ ,  $j_1, j_2 \in H$ .
- $\mathcal{E}_C$ :  $\min_{j \in |H|} |c_j| \geq \omega_3/(d^2 \log^2 d)$  for some  $\omega_3 > 0$ .



- $\mathcal{E}_{\hat{L}}: \|\Pi U^{\hat{L}}\|_1 \leq \omega_4/(d^2 \log^2 d)$  for some  $\omega_4 > 0$ .

Recall that we set  $s = \omega d^5 \log^5 d$  in Theorem 2. We will show that, with  $\omega$  sufficiently large and proper choices of  $\omega_1$ ,  $\omega_2$ ,  $\omega_3$ , and  $\omega_4$ , the event  $\mathcal{E}_U$  leads to an upper bound of  $\|\Pi y\|_1$  for all  $y \in \text{range}(A)$ ,  $\mathcal{E}_U$  and  $\mathcal{E}_L$  lead to a lower bound of  $\|\Pi y\|_1$  for all  $y \in Y^L$  with probability at least 0.9, and  $\mathcal{E}_H$ ,  $\mathcal{E}_{\hat{L}}$ , and  $\mathcal{E}_C$  together imply an lower bound of  $\|\Pi y\|_1$  for all  $y \in Y^H$ .

**Lemma 12.** *Provided  $\mathcal{E}_U$ , we have*

$$\|\Pi y\|_1 \leq \omega_1 d \log d \cdot \|y\|_1, \quad \forall y \in \text{range}(A).$$

*Proof.* For any  $y \in \text{range}(A)$ , we can find an  $x$  such that  $y = Ux$ . Then,

$$\|\Pi y\|_1 = \|\Pi Ux\|_1 \leq |\Pi U|_1 \|x\|_\infty \leq |\Pi U|_1 \|Ux\|_1 \leq \omega_1 d \log d \cdot \|y\|_1.$$

□

**Lemma 13.** *Provided  $\mathcal{E}_L$ , for any fixed  $y \in Y^L$ , we have*

$$\log \Pr \left[ \|\Pi y\|_1 \leq \frac{1}{4} \|y\|_1 \right] \leq -\frac{d \log d}{24\omega_2}.$$

*Proof.* Let  $z = \Pi y$ . We have,

$$|z_i| = \left| \sum_j s_{ij} c_j y_j \right| \simeq \left( \sum_j s_{ij} |y_j| \right) |\tilde{c}_i| \succeq \left( \sum_j s_{ij} |y_j^L| \right) |\tilde{c}_i| := \tilde{\gamma}_i |\tilde{c}_i|,$$

where  $\{\tilde{c}_i\}$  are independent Cauchy variables. Let  $\tilde{\gamma} = \sum_i \tilde{\gamma}_i = \|y^L\|_1$ . Since  $|y| \leq v$ , we have  $\tilde{\gamma}_i \leq \|Sv^L\|_\infty$ . By Lemma 6,

$$\log \Pr \left[ X \leq \frac{\|y^L\|_1}{2} \right] \leq \frac{-\|y^L\|_1}{12\|Sv^L\|_\infty}.$$

By assumption  $\mathcal{E}_L$  and  $\|y^L\|_1 \geq \frac{1}{2} \|y\|_1 \geq \frac{1}{2}$ , we obtain the result. □

**Lemma 14.** *Assume both  $\mathcal{E}_U$  and  $\mathcal{E}_L$ . If  $\omega_1$  and  $\omega_2$  satisfy*

$$d \log(6d(1 + 4\omega_1 d \log d)) - \frac{d \log d}{24\omega_2} \leq \log \delta$$

*for some  $\delta \in (0, 1)$  regardless of  $d$ , then, with probability at least  $1 - \delta$ , we have*

$$\|\Pi y\|_1 \geq \frac{1}{8} \|y\|_1, \quad \forall y \in Y^L.$$

*Proof.* Set  $\epsilon = 1/(2 + 8\omega_1 d \log d)$  and create an  $\epsilon$ -net  $Y_\epsilon^L \subseteq Y^L$  such that for any  $y \in Y^L$ , we can find a  $y_\epsilon \in Y_\epsilon^L$  such that  $\|y - y_\epsilon\|_1 \leq \epsilon$ . Since  $\|y\|_1 \leq d$  for all  $y \in Y^L$ , there exist such an  $\epsilon$ -net with at most  $(3d/\epsilon)^d$  elements (Bourgain et al. [4]). By Lemma 13, we can apply a union bound for all the elements in  $Y_\epsilon^L$ :

$$\Pr[\|\Pi y_\epsilon\|_1 \geq \frac{1}{4} \|y_\epsilon\|_1, \forall y_\epsilon \in Y_\epsilon^L] \geq 1 - \left( \frac{3d}{\epsilon} \right)^d e^{-\frac{d \log d}{24\omega_2}} = 1 - e^{d \log \frac{3d}{\epsilon} - \frac{d \log d}{24\omega_2}} \geq 1 - \delta.$$

For any  $y \in Y^L$ , we have, noting that  $y - y_\epsilon \in \text{range}(A)$ ,

$$\begin{aligned}\|\Pi y\|_1 &\geq \|\Pi y_\epsilon\|_1 - \|\Pi(y - y_\epsilon)\|_1 \geq \frac{1}{4}\|y_\epsilon\|_1 - \omega_1 d \log d \cdot \|y - y_\epsilon\|_1 \\ &\geq \frac{1}{4}\|y\|_1 - \left(\frac{1}{4} + \omega_1 d \log d\right) \epsilon \geq \frac{1}{8}\|y\|_1.\end{aligned}$$

So we establish a lower bound for all  $y \in Y^L$ . □

**Lemma 15.** *Provided  $\mathcal{E}_H$  and  $\mathcal{E}_{\hat{L}}$ , if  $\omega_3 > 4\omega_4$ , we have*

$$\|\Pi y\|_1 \geq \frac{\omega_4}{d^2 \log^2 d} \|y\|_1, \quad \forall y \in Y^H.$$

*Proof.* For any  $y = Ux \in Y^H$ , we have,

$$\begin{aligned}\|\Pi y\|_1 &\geq \|\Pi(y^H + y^{\hat{L}})\|_1 \geq \|\Pi y^H\|_1 - \|\Pi U^{\hat{L}} x\|_1, \\ &\geq \sum_{j \in H} |c_j| |y_j| - |\Pi U^{\hat{L}}|_1 \|x\|_\infty \geq \min_{j \in H} |c_j| \|y^H\|_1 - |\Pi U^{\hat{L}}|_1 \\ &\geq \left( \frac{\omega_3}{2d^2 \log^2 d} - \frac{\omega_4}{d^2 \log^2 d} \right) \|y\|_1 \geq \frac{\omega_4}{d^2 \log^2 d} \cdot \|y\|_1,\end{aligned}$$

which creates a lower bound for all  $y \in Y^H$ . □

We continue to show that, with  $\omega$  sufficiently large, by setting  $\tau = \omega^{1/4}/(d \log^2 d)$  and choosing  $\omega_1, \omega_2, \omega_3$ , and  $\omega_4$  properly, we have each event with probability at least  $1 - 0.08 = 0.92$  and thus

$$\Pr[\mathcal{E}_U \cap \mathcal{E}_L \cap \mathcal{E}_H \cap \mathcal{E}_{\hat{L}} \cap \mathcal{E}_C] \geq 0.6.$$

Moreover, the condition in Lemma 14 holds with  $\delta = 0.1$ , and the condition in Lemma 15 holds. Therefore,  $\Pi = SC$  has the desired property with probability at least 0.5, which would conclude the proof of Theorem 2.

**Lemma 16.** *With probability at least 0.92,  $\mathcal{E}_U$  holds with  $\omega_1 = 500(1 + \log \omega)$ .*

*Proof.* With  $S$  fixed, we have,

$$|\Pi U|_1 = |SCU|_1 = \sum_{k=1}^d \sum_{i=1}^s \left| \sum_{j=1}^n s_{ij} c_j u_{jk} \right| \simeq \sum_{k=1}^d \sum_{i=1}^s \sum_{j=1}^n (|s_{ij} u_{jk}|) |\tilde{c}_{ik}|,$$

where  $\{\tilde{c}_{ik}\}$  are *dependent* Cauchy random variables. We have

$$\sum_{k=1}^d \sum_{i=1}^s \sum_{j=1}^n |s_{ij} u_{jk}| = \sum_{k=1}^d \sum_{j=1}^n |u_{jk}| = |U|_1 = d.$$

Apply Lemma 5,

$$\Pr[|\Pi U|_1 \geq td \mid S] \leq \frac{2 \log(sdt)}{t}.$$

Setting  $\omega_1 = 500(1 + \log \omega)$  and  $t = \omega_1 \log d$ , we have

$$\frac{2 \log(sdt)}{t} = \frac{2 \log(\omega \omega_1 d^6 \log^5 d)}{\omega_1 \log d} \leq 0.08.$$

We assume that  $\log d \geq 1$  and  $\log \omega \geq 1$ . □

**Lemma 17.** For any  $\delta \in (0, 0.1)$ , if  $s \geq d/\tau$ , we have,

$$\Pr \left[ \|Sv^L\|_\infty \geq \left(1 + 2 \log \frac{d}{\delta\tau}\right) \cdot \tau \right] \leq \delta.$$

*Proof.* Let  $X_{ij} = s_{ij}v_j^L$ . We have  $\mathbf{E}[X_{ij}] = v_j^L/s$ ,  $\mathbf{E}[X_{ij}^2] = (v_j^L)^2/s$ , and  $0 \leq X_{ij} \leq v_j^L \leq \tau$ . Fixed  $i$ ,  $X_{ij}$  are independent,  $j = 1, \dots, n$ . By Bernstein's inequality,

$$\log \Pr \left[ \sum_j X_{ij} \geq \frac{\|v^L\|_1}{s} + t \right] \leq \frac{-t^2/2}{\|v^L\|_2^2/s + \tau t/3} \leq \frac{-t^2/2}{\tau(\|v^L\|_1/s + t/3)} \leq \frac{-t^2/(2\tau)}{d/s + t/3}.$$

where we use Holder's inequality:  $\|v^L\|_2^2 \leq \|v^L\|_1 \|v^L\|_\infty \leq d\tau$ . To obtain a union bound for all  $i$  with probability  $1 - \delta$ , we need

$$\frac{-t^2/(2\tau)}{d/s + t/3} + \log s \leq \log \delta.$$

Given  $\delta < 0.1$ , it suffices to choose  $s = d/\tau$  and  $t = 2 \log(d/(\delta\tau))\tau$ . Note that  $\|v^L\|_1/s \leq \|v\|_1/s = \tau$ . We have

$$\Pr \left[ \|Sv^L\|_\infty \geq \left(1 + 2 \log \frac{d}{\delta\tau}\right) \cdot \tau \right] \leq \delta.$$

Increasing  $s$  will decrease the failure rate, so it holds for all  $s \geq d/\tau$ .  $\square$

**Lemma 18.** With probability at least 0.92,  $\mathcal{E}_L$  holds with  $\omega_2 = (15 + \log \omega)/\omega^{1/4}$ .

*Proof.* By Lemma 17, with probability at least 0.92,  $\mathcal{E}_L$  holds with

$$\omega_2 = \frac{1 + 2 \log \frac{\omega^{1/4} d^2 \log^2 d}{0.08}}{\omega^{1/4} \log d} \leq \frac{15 + \log \omega}{\omega^{1/4}}.$$

$\square$

**Lemma 19.** With the above choices of  $\omega_1$  and  $\omega_2$ , the condition in Lemma 13 holds with  $\delta = 0.1$  for sufficiently large  $\omega$ .

*Proof.* With  $\omega_1 = 500(1 + \log \omega)$ , and  $\omega_2 = (15 + \log \omega)/\omega^{1/4}$ , the first term in

$$d \log(6d(1 + 4\omega_1 d \log d)) - \frac{d \log d}{24\omega_2}$$

increases much slower than the second term as  $\omega$  increases, while both are at the order of  $d \log d$ . Therefore, if  $\omega$  is sufficiently large, the condition hold with  $\delta = 0.1$ .  $\square$

**Lemma 20.** If  $\omega \geq 160$ , event  $\mathcal{E}_H$  holds with probability at least 0.92.

*Proof.* Given  $j_1, j_2 \in H$  and  $j_1 \neq j_2$ , let  $X_{j_1 j_2} = 1$  if  $\phi(j_1) = \phi(j_2)$  and  $X_{j_1 j_2} = 0$  otherwise. It is easy to see that  $\Pr[X_{j_1 j_2} = 1] = \frac{1}{s}$ . Therefore,

$$\Pr[\mathcal{E}_H] \geq 1 - \sum_{j_1 < j_2} \Pr[X_{j_1 j_2} = 1] \geq 1 - \frac{|H|^2}{s} \geq 1 - \frac{d^2}{s\tau^2} \geq 1 - \frac{1}{\omega^{1/2}}.$$

It suffices if  $\omega \geq 160$ .  $\square$

**Lemma 21.** *With probability at least 0.92, event  $\mathcal{E}_C$  holds with  $\omega_3 = 1/(8\omega^{1/4})$ .*

*Proof.* Let  $c$  be a Cauchy variable. We have

$$\Pr[|c| \leq t] = \frac{2}{\pi} \tan^{-1} t \leq \frac{2t}{\pi}.$$

$|H|$  is at most  $d/\tau = \omega^{1/4} d^2 \log^2 d$ . Then

$$\begin{aligned} \Pr[\mathcal{E}_C] &\geq 1 - |H| \cdot \Pr\left[|c| < \frac{\omega_3}{d^2 \log^2 d}\right] \\ &\geq 1 - \omega^{1/4} d^2 \log^2 d \cdot \frac{2\omega_3}{\pi d^2 \log^2 d}. \end{aligned}$$

Therefore,  $\omega_3 = 1/(8\omega^{1/4})$  would suffice.  $\square$

**Lemma 22.** *With probability at least 0.92, event  $\mathcal{E}_{\hat{L}}$  holds with  $\omega_4 = 25000(1 + \log \omega)/\omega^{3/4}$ . Thus with  $\omega$  sufficiently large and the above choice of  $\omega_3$ , the condition in Lemma 15  $\omega_3 > 4\omega_4$  holds.*

*Proof.* We have,

$$\mathbf{E}[|U^{\hat{L}}|_1] = \frac{|H|}{s} |U^L|_1 \leq \frac{\omega^{1/4} d^2 \log^2 d}{\omega d^5 \log^5 d} \cdot d = \frac{1}{\omega^{3/4} d^2 \log^3 d}.$$

By Markov's inequality,

$$\Pr\left[|U^{\hat{L}}|_1 \geq \frac{25}{\omega^{3/4} d^2 \log^3 d}\right] \leq 0.04.$$

Assume that  $|U^{\hat{L}}|_1 \leq \frac{25}{\omega^{3/4} d^2 \log^3 d}$ . Similar to the proof of Lemma 16, we have

$$|\Pi U^{\hat{L}}|_1 = \sum_{k=1}^d \sum_{i \in \phi(H)} \left| \sum_j s_{ij} c_j u_{jk}^{\hat{L}} \right| \simeq \sum_{k=1}^d \sum_{i \in \phi(H)} \left( \sum_j s_{ij} |u_{jk}^{\hat{L}}| \right) |\tilde{c}_{ik}|,$$

where  $\{\tilde{c}_{ik}\}$  are *dependent* Cauchy variables. Apply Lemma 5,

$$\Pr[|\Pi U^{\hat{L}}| \geq |U^{\hat{L}}|t] \leq \frac{2 \log(|H|dt)}{t}$$

It suffices to choose  $t = 1000(1 + \log \omega) \log d$  to make the RHS less than 0.04. So with probability at least 0.92, we have  $\mathcal{E}_{\hat{L}}$  holds with  $\omega_4 = 25000(1 + \log \omega)/\omega^{3/4}$ .  $\square$

### A.3 Proof of Corollary 2 (Fast $\ell_1$ Regression)

By Theorem 2 and Lemma 3, we know that Steps 2 and 4 of Algorithm 1 succeed with a constant probability. Conditioning on this event, we have

$$\|A\hat{x} - b\|_1 \leq \frac{1}{1 - \epsilon/4} \|SA\hat{x} - Sb\|_1 \leq \frac{1 + \epsilon/4}{1 - \epsilon/4} \|SAx^* - Sb\|_1 \leq \frac{(1 + \epsilon/4)^2}{1 - \epsilon/4} \|Ax^* - b\|_1 \leq (1 + \epsilon) \|Ax^* - b\|_1,$$

where the last inequality is due to  $\epsilon < 1/2$ . By Theorem 2, Step 2 takes  $\mathcal{O}(\text{nnz}(A))$  time, and Step 3 takes  $\mathcal{O}(\text{poly}(d))$  time because  $\Pi A$  has  $\mathcal{O}(\text{poly}(d))$  rows. Then, by Lemma 3, Step 4 takes  $\mathcal{O}(\text{nnz}(A) \cdot \log n)$  time, and Step 5 takes  $\mathcal{T}_1(\epsilon/4; \mathcal{O}(\text{poly}(d) \log(1/\epsilon)/\epsilon^2), d)$  time. Therefore, the total running time of Algorithm 1 is as stated.

#### A.4 Proof of Lemma 8

First, we know that

$$\Pr[|X_p|^p \geq t] = \Pr[|X_p| \geq t^{1/p}] = 2 \cdot \Pr[X_p \geq t^{1/p}].$$

Next, we state the following lemma, which is due to Nolan [25].

**Lemma 23.** (Nolan [25, Thm. 1.12]) *Let  $X \sim \mathcal{D}_p$  with  $p \in [1, 2)$ . Then as  $x \rightarrow \infty$ ,*

$$\Pr[X > x] \sim c_p x^{-p},$$

where  $c_p = \sin \frac{\pi p}{2} \cdot \Gamma(p)/\pi$ .

By Lemma 23, it follows that, as  $t \rightarrow \infty$ ,

$$\Pr[|X_p|^p \geq t] \sim 2c_p t^{-1}.$$

For the Cauchy distribution, we have

$$\Pr[|C| \geq t] = 1 - \frac{2}{\pi} \tan^{-1} t = \frac{2}{\pi} \tan^{-1} \frac{1}{t} \sim \frac{2}{\pi} \cdot t^{-1}.$$

Hence, there exist  $\alpha'_p > 0$  and  $t_1 > 0$  such that for all  $t > t_1$ ,

$$\Pr[\alpha'_p |C| \geq t] \geq \Pr[|X_p|^p \geq t].$$

Note that all the  $p$ -stable distributions with  $p \in [1, 2]$  have finite and positive density at  $x = 0$ . Therefore, there exists  $\alpha''_p > 0$  such that for all  $0 \leq t \leq t_1$ ,

$$\Pr[\alpha''_p |C| \geq t] \geq \Pr[|X_p|^p \geq t].$$

Let  $\alpha_p = \max\{\alpha'_p, \alpha''_p\}$ . We get  $\alpha_p |C| \succeq |X_p|^p$ . For the Gaussian distribution, we have, as  $t \rightarrow \infty$ ,

$$\Pr[|G|^2 \geq t] \sim 2e^{-t/2} t^{-1/2}.$$

which converges to zero much faster than  $t^{-1}$ , so we can apply similar arguments to obtain  $\beta_p$ .

#### A.5 Proof of Lemma 9 (Upper Tail Inequality for $p$ -stable Distributions)

Let  $C_i = F_c^{-1}(F_p(X_i))$ ,  $i = 1, \dots, m$ , where  $F_c$  is the CDF of the standard Cauchy distribution and  $F_p$  is the CDF of  $\mathcal{D}_p$ .  $C_i$  follows the standard Cauchy distribution, and, by Lemma 8, we have  $\alpha_p |C_i| \geq |X_i|^p$ . Therefore, for any  $t \geq 1$ ,

$$\Pr[X \geq t\alpha_p \gamma] \leq \Pr\left[\sum_i \gamma_i |C_i| \geq t\gamma\right] \leq \frac{2 \log(mt)}{t}.$$

The last inequality is from Lemma 5.

#### A.6 Proof of Lemma 10 (Lower Tail Inequality for $p$ -stable Distributions)

Let  $G_i$  be independent random variables sampled from the standard Gaussian distribution,  $i = 1, \dots, m$ . By Lemma 8, we have

$$\log \Pr[X \leq \beta_p (1-t)\gamma] \leq \log \Pr\left[\sum_i \gamma_i |G_i|^2 \leq (1-t)\gamma\right].$$

The lower tail inequality from Lemma 7 concludes the proof.

### A.7 Proof of Theorem 6 (Low-distortion Dense Embedding for $\ell_p$ )

The proof is similar to the proof of Sohler and Woodruff [28, Theorem 5], except that the Cauchy tail inequalities are replaced by tail inequalities for the stable distributions. For simplicity, we omit the complete proof but show where to apply those tail inequalities. By Lemma 11, there exists a  $(d^{1/p}, 1, p)$ -conditioned basis matrix of  $\mathcal{A}_p$ , denoted by  $U$ . Thus,  $|U|_p^p = d$ , where recall that  $|\cdot|_p$  denotes the element-wise  $\ell_p$  norm of a matrix. We have,

$$|\Pi U|_p^p = \sum_{k=1}^d \|\Pi u_k\|_p^p = \sum_{k=1}^d \sum_{i=1}^s \left| \sum_{j=1}^n \Pi_{ij} u_{jk} \right|^p \simeq \sum_{k=1}^d \sum_{i=1}^s \|u_k\|_p^p |\tilde{X}_{ik}|^p,$$

where  $\tilde{X}_{ik} \sim \mathcal{D}_p$ . Applying Lemma 9, we get  $\|\Pi U\|_p^p/s = \mathcal{O}(d \log d)$  with a constant probability. Define  $Y = \{Ux \mid \|x\|_q = 1, x \in \mathbb{R}^d\}$ . For any fixed  $y \in Y$ , we have

$$\|\Pi y\|_p^p = \sum_{i=1}^s \left| \sum_{j=1}^n \Pi_{ij} y_j \right|^p \simeq \sum_{i=1}^s \|y\|_p^p |\tilde{X}_i|^p,$$

where  $\tilde{X}_i \stackrel{\text{iid}}{\sim} \mathcal{D}_p$ . Applying Lemma 9, we get  $\|\Pi y\|_p^p/s \leq 1/\mathcal{O}(1)$  with an exponentially small probability with respect to  $s$ . By choosing  $s = \omega d \log d$  with  $\omega$  sufficiently large and an  $\epsilon$ -net argument on  $Y$ , we can obtain a union lower bound of  $\|\Pi y\|_p^p$  on all the elements of  $Y$  with a constant probability. Then,

$$1/\mathcal{O}(1) \cdot \|y\|_p^p \leq \|\Pi y\|_p^p/s \leq |\Pi U|_p^p \|x\|_q^p \leq \mathcal{O}(d \log d) \cdot \|Ux\|_p^p = \mathcal{O}(d \log d) \|y\|_p^p, \quad y \in Y,$$

which gives us the desired result.

### A.8 Proof of Theorem 7 (Improving the Embedding Dimension)

Each of Steps 1, 3, and 5 of Algorithm 2 succeeds with a constant probability and we can control the failure rate by repeating trials. Thus, with a constant probability, all steps succeed. Conditioning on this event, we have  $\kappa_p(AR^{-1}) = 6d$  because

$$\begin{aligned} \|AR^{-1}x\|_p &\leq 2\|\tilde{S}AR^{-1}x\|_p \leq 4d\|x\|_2, \\ \|AR^{-1}x\|_p &\geq \frac{2}{3}\|\tilde{S}AR^{-1}x\|_p \geq \frac{2}{3}\|x\|_2, \quad \forall x \in \mathbb{R}^d. \end{aligned}$$

By Lemma 1,  $\bar{\kappa}_p(AR^{-1}) \leq 6d^{1/p+1}$ , and then by Lemma 3, the embedding dimension of  $S$  is  $\mathcal{O}(\bar{\kappa}_p^p(AR^{-1})d^{\lfloor p/2-1 \rfloor} d \log(1/\epsilon)/\epsilon^2) = \mathcal{O}(d^{3+p/2} \log(1/\epsilon)/\epsilon^2)$ .